Preprint from: "PTB-Bericht", Series IT (Information Technology), 2003-02-10
PTB Workshop on uncertainty, December 2002

Classes of inter-comparisons and the case of temperature standards

Franco Pavese

Istituto di Metrologia «G.Colonnetti», CNR, Torino, Italy, f.pavese@imgc.cnr.it

Patrizia Ciarlini

Istituto per le Applicazioni del Calcolo «M.Picone», CNR, Roma, Italy, ciarlini@iac.rm.cnr.it

ABSTRACT

The paper discussed a classification of the inter-comparisons that is relevant to identify the proper statistical method to combine the data provided for each participant to the inter-comparison.
The proposed approach for Class 2 IC's constructs a single probabilistic model for the reference distribution probability function, based on the use of the mixture density model. This approach allows the estimate of the reference value simply as the expected value of the mixture density function. The method does not require strong assumptions –as *N* large– or limitations on the local probability distributions, such as the Normality density in each participant laboratory. It is particularly valuable for inter-comparisons of physical-state realisations, e.g., for temperature standards, where the population associated to an IC can be viewed as a super-population. However, Class 2 inter-comparisons probably include a wider range of IC's, such as the ones where a single standard is circulated and measured. The paper also compared the main features of the statistical treatments suitable for the IC outcomes in both cases of Class 1 and Class 2. Some aspects of the treated problems, especially the meaning of uncertainty in Class 2 IC's and the consequent approach, still deserve a deeper subsequent insight.

1. INTRODUCTION

Inter-laboratory comparisons (inter-comparisons, IC's) are an essential excercise for evaluating the international equivalence between the National Metrology Institutes (NMI's).
There are some basic questions that should find the correct answers before the exercises take place, which often remain hidden in the Protocols, such as:

1. Does every type of inter-comparisons require the same type of statistical treatment;
2. Can the data from each participant always be considered as sampled from a separate stochastic variable, or are there cases where they are samples from a single population (single stochastic variable);
3. When should the data from each participant to an inter-comparison be treated as *pooled* and when as *hierarchical* data;
4. Which is the minimum information supplied for each participant that should be considered sufficient for the subsequent statistical treatment;
5. Can a Reference Value always accurately summarise the inter-comparison results;
6. Is the choice of attributing or not an uncertainty to the Reference Value a free one, i.e., can it sometimes be correctly considered as a deterministic parameter
7. Which is the best method for evaluating the inter-comparison uncertainty.

Other peculiarities in most inter-comparisons, such as the small size of the samples to be analysed, which corresponds to the number of participants, or the attitude to include every participant results, should be taken into account in determining the procedures to adopt in the IC computations. Recent papers [1–5] introduced first attempts to provide tools for some of the answers. In particular the use in IC's of the mixture probability distribution was first proposed in [4].
This paper focuses on a discussion about the need to adapt the statistical treatment of the inter-comparison data to the characteristics of different inter-comparisons, starting from the definition of a classification of the IC's first suggested in [5]. Several consequences for the different treatments that can be applied will be also underlined.

The paper discusses the previous issues with the aim to interpret in an appropriate statistical framework the assumptions underlying the analysis of the key comparison data, when no correlation exists between participants. The paper concentrates on the case of temperature standards, as an example of a peculiar class of inter-comparisons.

2. CLASSES OF INTERCOMPARISONS

Most inter-comparisons concern standards consisting of local artefacts. Each artefact carries its own value of the relevant quantity, which can be estimated only by calculation or/and through comparison with other standards of similar type. In other words, normally, there is not a «natural» value available to characterise these standards. Examples are the standards for mass or for length (gauge blocks). For these inter-comparisons, as discussed in [5], the input data to the IC for each participant to the inter-comparison should be suitably combined to output a single value and treated as *hierarchical* data.

However, there is another type of inter-comparisons, whose aim is to compare independent local realisations of a given physical state[1]. Can they be included in the previous class ?

Let us take, as a study case, the standards based on temperature fixed-points as defined by the ITS-90: values of the electrical resistance of each participant thermometer (a pure-platinum wire, SPRT), are experimentally determined at the temperature of each of a number of physical states. The realisation of each physical state does not depend, for the triple points, from any other thermodynamic quantity, nor depends on the temperature value assigned to the state. A fixed point of this kind is a purely thermodynamic experiment (phase transition, whose temperature value T is unique in Nature), whose practical implementation is performed according to the state-of-the-art practice, thus providing an experimental approximation of the «natural» temperature value T.

Figure 1 [3] shows the meaning of the implementation in each NMI.

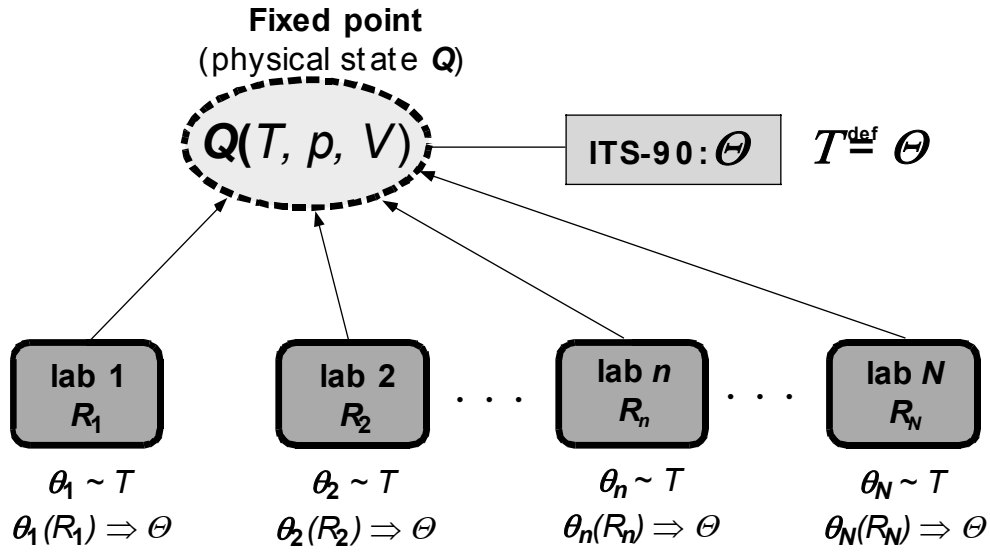

Fig.1. Laboratory independent realisations of the physical quantity Q (fixed point).

Each n-th Laboratory assigns to its realisation the same temperature value $\theta_n \Rightarrow \Theta$, the one defined by the ITS-90, to which corresponds the value R_n of its SPRT thermometer.

[1] The physical-state definition is assumed to be unambiguous, at least to such a level that no significant contribution to the total uncertainty of the standards arises.

The *n*-th participant can obtain evidence that the measured R_n corresponds to the correct temperature of the given phase transition (the physical state in question, Q) by comparing its own realisation with different realisations and different practical implementations of the same physical state provided by other Laboratories, through the thermometers –i.e., through the R_j values.

The inter-comparison shows (Fig. 2 [3]) that, in fact, the θ_n in Fig.1 are different each other, i.e., that most $\Delta\theta_{0n} \neq 0$. Consequently, only one –or, most probably, none– of the Laboratories realises Θ exactly (in fact within the Laboratory uncertainty) [2]: the differences between θ_n and Θ remain *un-revealed* by the inter-comparison. One of the consequences is that it is not possible to model the individual biases with respect to □. In the following, the idea is not to identify a statistical procedure able to compensate for these possible biases, but, instead, to consider them as peculiarities of the given IC that must be taken into account in a single statistical model.

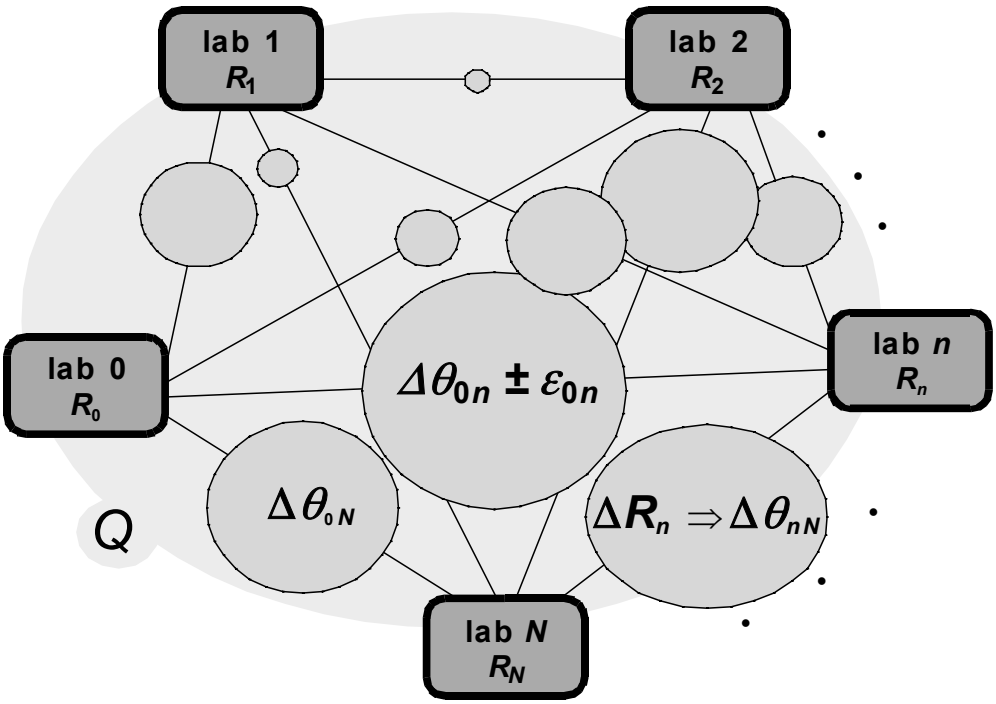

Fig.2. Inter-comparison of independent Laboratory realisations at the *same* physical state Q. Thermometer-response differences ΔR_{jk} are measured: they can be transformed in $\Delta\theta_{jk}$ values simply through the known sensitivity dR/dT of each thermometer.

In any instance, by no way these possible biases between Laboratory realisations, i.e. the fact that different estimates of the temperature value of the physical state can be different, affects the fact

[2] The temperature differences in Fig.2 can be transformed in temperature differences only by *arbitrarily* assuming one Laboratory as the «reference» one (e.g., Lab 0), and assigning the ITS-90 temperature value Θ to its realisation ($\theta_0 \equiv \Theta$): then the θ_k values for each *k*-th Laboratory come. From the set $\{\theta_k\}$ one can define a single summary value $\theta_r = f(\theta_k)$ representing the comparison «reference value». The difference of θ_r from Θ (or T) remains unknown, not being an IC outcome.

that the inter-comparison measurand is a single physical state and that the latter carries a unique *T* value [3].

Each Laboratory aims at realising the true physical state, not using a «reference material» or an artefact. Consequently, all practical devices for realising each fixed point should be considered as sampling from the same population, being all experimental realisations of the same physical state [3]. The previous considerations are equivalent to assess that the inter-comparison data have a homogeneous nature in some sense.

In conclusion, the Laboratory samples can be viewed as a population pertaining to the super-population represented by the inter-comparison. In this case, the statistical analysis should apply to a super-population formed by *the union of* the individual Laboratory populations [5]. In other words, the total sample of the inter-comparison can be viewed as the one obtained by *pooling* the samples of the participants

Consequently, two main types of inter-comparisons (and of standards) pertaining to two distinct classes can be defined: 1) artefact inter-comparisons, 2) physical-state realisation inter-comparisons. The measurand for the two classes has a different statistical meaning and deserve the choice of a different statistical treatment.

3. THE INTERCOMPARISON DATA FOR EACH PARTICIPANT

The probabilistic model of the IC data for each of *N* participants is normally summarised by three items, representing the statistical properties of the local standard:

1. $\pm U$, the extended point uncertainty, generally set at a 95% confidence level;
2. Y, the value assigned to the local standard and, in a probabilistic sense, the location parameter of the local population;
3. F, the statistical distribution, associated to the local standard to identify the population, which GUM also requires to be supplied.

Thus, the IC data in the following form is supposed to completely represent the probabilistic and metrological information concerning every *j*-th participant Laboratory:

$$[(Y_j \pm U_j),\ F_j] \qquad j=1,\ldots,N \tag{1}$$

It assumes that the underlying density distribution function is symmetric and that is centred on the location parameter. [4]

3.1 Data of Artefact Inter-comparisons (Class 1)

Each artefact –i.e., each local standard– is a distinct individual, which can be stochastically modelled. Hence, the sample $(x_{j1},\ldots, x_{jH_j})$ of the *j*-th participant[5], pertain to a *distinct random variable*, X_j

$$x_{ji} \sim X_j \qquad j = 1,\ldots, N\,. \tag{2}$$

In this case, the items of the probabilistic model (1) have the following meaning:

[3] Scale-related non-uniqueness is a matter of different origin and nature and is irrelevant to the contents of this paper.

[4] It must be noted that in temperature metrology skewed densities can occur.

[5] The local samples most often are not made available to the inter-comparison.

- Y_j: $Y_j = S(X_j)$, obtained by choosing a suitable statistics S, whose estimate $y_j = S(x_{j1}, \ldots, x_{jH_j})$ aims to approximate the expected value $E(X_j)$ with respect to the local pdf; [6]
- U_j: is the extended uncertainty associated to $S(X_j)$, the summary value;
- F_j: actually, Y_j depends on the local distribution F_j , i.e., $Y_j = S(X_j, F_j)$.

Usually the sizes H_j of the local samples are not provided. It must be reminded that a small size of the samples may not provide credible estimates from the participant. In this Class of IC's it must be noted that, the homogeneity in the data (1) may be not verified. Moreover, the statistical inference on the IC data concerns the analysis of the IC sample $(y_1, \ldots, y_N)$ in the frequency approach and the items y_j have a *hierarchical nature*, actually y_j is obtained by estimating a summary statistics for the corresponding local sample $(x_{j1}, \ldots, x_{jH_j})$.

3.2 Data of Physical-state Realisation Inter-comparisons (Class 2)

Each participant aims at locally approximating the very same physical state Q –e.g., the fixed-point temperature in the example of Section 2. The IC measurand is *the* physical state, as shown in Fig. 1, and all values measured from independent experimental realisation of that physical state should be considered as sampling from a single pdf, i.e., from a *single stochastic variable, say Q*, i.e. from a single pdf. Hence, for the samples $(x_{j1}, \ldots, x_{jh}, \ldots, x_{jH_j})$ of the j-th participant [4]

$$x_{jh} \sim Q \qquad \forall\, j\,. \tag{3}$$

Often the size of the sample, H_i, acquired in some Laboratories to infer the required statistical information Y_j and U_j is very small from a statistical point of view, mainly due to the high cost of the experiments.
Here the items in (1) have the following meaning:

- F_j: is the local probability distribution of the j-th local population. At least a general class of models for it should be provided (for example, the normal, the uniform and the triangular are classes of distribution functions discussed in [4] for cases in thermal metrology).
- Y_j, and U_j: the definition of a new stochastic variable Y_j that combines the local sample to estimate a location parameter and its associated uncertainty –i.e., the local summary parameters– should be avoided, since all local datasets $(x_{j1}, \ldots, x_{jH_j})$ pertain to the same population, being samples from the same stochastic variable Q . Often, the provided U_j actually has the broader meaning of representing the whole *a priori* information contained in the Laboratory uncertainty budget, especially when type-B errors are dominant.[7]

In this framework, the overall data are considered as a homogeneous population. So doing, if all the local pdf's were identical, the output pdf would be exactly the same too. Hence, it can be appropriate to *pool* the local samples into a single sample (of total size $N \cdot \sum(H_j)$) and to estimate the mean value from it. This is equivalent to assume that a single distribution function, called the *compound distribution*, stochastically describing the variability of the pooled sample is the correct probabilistic model for data that do not have a hierarchical nature. Summary operations performed in each Laboratory should be avoided, since they could introduce an improper hierarchical variability and the statistical treatment should not concern the framework of repeated measurements analysis.

[6] The expected value of the RV, X, or the mean value $m = \int x\, f(x)\, dx$, which can be different from the estimate y_j.
[7] This meaning of U_j often also apply to Class 1 IC's.

4. INTERCOMPARISON OUTCOMES: PAIR DEGREE OF EQUIVALENCE

The Pair Degree of Equivalence (PDE), defined as the difference between the measurand values of every pair of participant Laboratories is the direct IC outcome, which does not need to resort to a Reference Value. It can directly be obtained from the IC results.
Let us consider the pair of the *n*-th and the *m*-th participant, with $n, m = 1...N$, $\forall\ n \neq m$.

4.1 PDE for Artefact Inter-comparisons

The pair degree of equivalence, $Y_{n,m}$, is a *new stochastic variable*, of *second-rank hierarchical level*:

$$Y_{n,m} = Y_n - Y_m \tag{4a}$$
$$F_{n,m} = F_n * F_m \tag{4b}$$
$$u_{n,m} = u(y_{n,m}) \tag{4c}$$

The resulting pdf is the convolution of F_n and F_m. When Y_n and Y_m are independent random variables having distributions $F_n = N(\mu_n, \sigma_n^2)$ and $F_m = N(\mu_m, \sigma_m^2)$, then the distribution of the combination (5a) is $F_{n,m} = N(\mu_n - \mu_m, \sigma_n^2 + \sigma_m^2)$.

Most often, the IC protocol does not specify all details for the local statistical treatments: for example, the minimum size of the local samples that has to be statistically analysed in the sense of the repeated measures is not provided. In addition, the treatment performed to estimate y_n and y_m could be different from participant to participant, bringing into $y_{n,m}$ possible biases. Should it happen that the *m*-th participant chooses the median as the summary statistics, while the *n*-th participant chooses the mean, should be taken into account when computing (4).

4.2 PDE for Physical-state Realisation Inter-comparisons

Two different statistical treatments are possible, depending on the type of information available for each participant Laboratory.

a) The original samples from each j-th population (random variable) are provided:

$$\{x_{ni}\} \text{ and } \{x_{mh}\}\ , \qquad i = 1,...H_n\ , \qquad h = 1,...H_m, \quad \forall\ n \neq m\ , \tag{5}$$

where every $x_{\bullet}$ is a sample of size H_n or H_m , respectively, from the same stochastic variable Q. In order to avoid the use of summary variables and the sampling theory when the size of the local sample is too small, the most straightforward way to estimate the difference of the two populations seems, in this case, the use of a resampling method, for example the bootstrap [2, 6], which estimates the difference of means only on the basis of the two given samples. Simulated differences are obtained by choosing at random one item, say x_{nk}, from the first sample $\{x_{ni}\}$, and by defining $x'_{ni} = x_{nk}$; by repeating H_n times to simulate a bootstrap sample $\{x'_{ni}\}$ of size H_n (the same procedure is applied to the second sample to generate $\{x'_{mh}\}$); by computing the means of the simulated samples and their difference $y'_{n,m} = \bar{x}'_n - \bar{x}'_{m,}$. This operation is repeated B times (with B large), the simulated differences are used to approximate the sampling distribution of $Y_{n,m}$, and the histogram is built via the B replications $y'_{n,m,b}$. Then the bootstrap estimate of the expected value and the related bootstrap interval at 95% of confidence level can be computed by applying the bootstrap estimates (4) to identify the pair degree of equivalence.

b) The original samples of each population are not provided:

F_n and F_m are only available.

In this case, one should resort to the density functions f_n and f_m of the two participants and compute from them in terms of the difference of their expected values, say $d_{n,m}$. Then, the value of the PDE is defined to be:

$$d_{n,m} = E(F_n) - E(F_m) \quad , \tag{6}$$

where $d_{n,m}$ is not a random variable but a *numerical value.* It cannot be interpreted as an estimate or an approximation of some unknown value. The computation in (6) can be performed in closed form; however, a Monte Carlo technique can be used as an alternative approach [6–8], if some density functions are not available in analytical form.

5. A STATISTICAL MODEL FOR CLASS 2 INTERCOMPARISONS

Summarising, for Class 2 inter-comparisons the following description of the IC data applies:

- All samples $\{\boldsymbol{x}_{ji}\}_{j=1,\ldots,N}$ are samples from Q
- Q is distributed according to a single (compound) distribution F, while each F_j represents the stochastic variability of the j-th local population.

In [4] a suitable probability model, able to represent the stochastic variability of composite populations, has been introduced. It is the combination of several *density functions, or mixture density model* [9]. It is adopted, for example, to model the distribution of quality data in industrial productions as a super-population: a stockpile of items made up of two components with two different lifetimes is identified by a mixture of two exponential distributions. Mixture distributions are also used to model important experimental situations in non-normal cases. Mixtures of normal densities, also known as "contaminated" normal family, are used to model a population, which follows a normal distribution except in those occasions where a peculiar observation is recorded. These peculiar observations are not viewed as outliers, but as the effects of an admissible variability in the IC measurements.

The mixture density function has the following general mathematical form:

$$f(x, \boldsymbol{\Lambda}) = \sum_{j=1}^{N} \pi_j f_j(\boldsymbol{x}; \boldsymbol{\Lambda}_j) \, , \tag{7a}$$

where $f_j(\boldsymbol{x};\boldsymbol{\Lambda}_j)$ is the density function of the distribution function F_j.
The mixture function depends on the parameter vector $\boldsymbol{\Lambda} = (\pi; \Lambda_1, \ldots, \Lambda_N$ and $\pi_j \geq 0$ are *proportion parameters*,

$$\sum_{j=1}^{N} \pi_j = 1, j = 1, \ldots, N \, . \tag{7b}$$

This finite mixture density function is a linear superposition of the N local densities, where each function $f_j(x, \boldsymbol{\Lambda}_j)$ is the local density associated to each Laboratory. Let $F(x,\boldsymbol{\Lambda}_j)$ be the so-called *compound distribution* corresponding to the mixture density in (7a). Under this probability model

each observation x_i is viewed as arising from a super-population that is a mixture of N populations in some proportion. It directly describes the total variability of the super-population (in IC literature also called *pooled population* [10]), as a whole. In this respect, the mixture function describes the variability of the super-population of the IC, as it could happen for the population related to a «super-laboratory». By assuming $1/\pi_j = 1/N$ for each j, it is supposed that each participant contributes to the IC in the same proportion, or, equivalently, with a same fraction of local data to form the pooled sample ($x_1, \ldots x_{N\Sigma H_j}$), where they are realised values independently and identically distributed with common distribution $F(\boldsymbol{x}, \boldsymbol{\Lambda})$.

In Figs.3 and 4 two examples of mixture density distributions with equal proportions are reported, which have been generated assuming a specific Normal density for each Laboratory to construct the histogram related to the key comparisons CCT K2 ($N = 7$) and CCT K3 ($N = 13$), respectively [11–12].

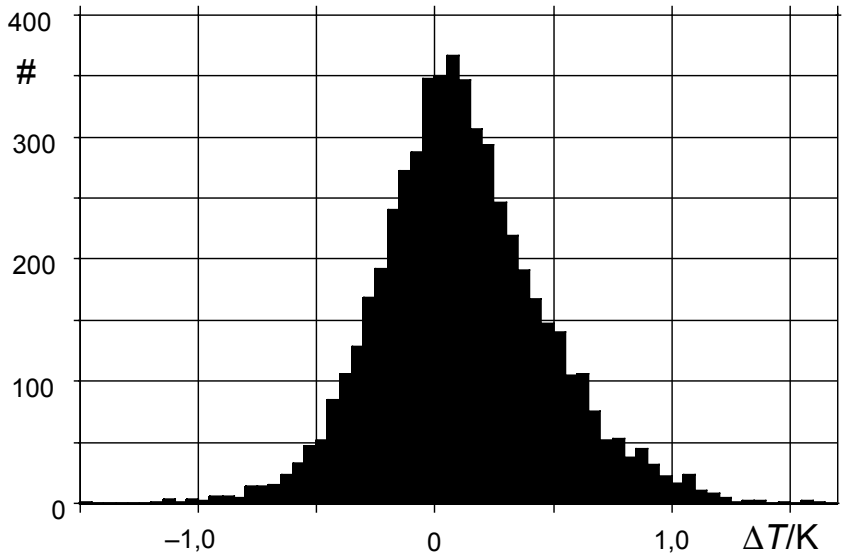

Fig.3. Mixture distributions for the results of CCT key comparisons: K2, triple point of argon ($B = 5500$).

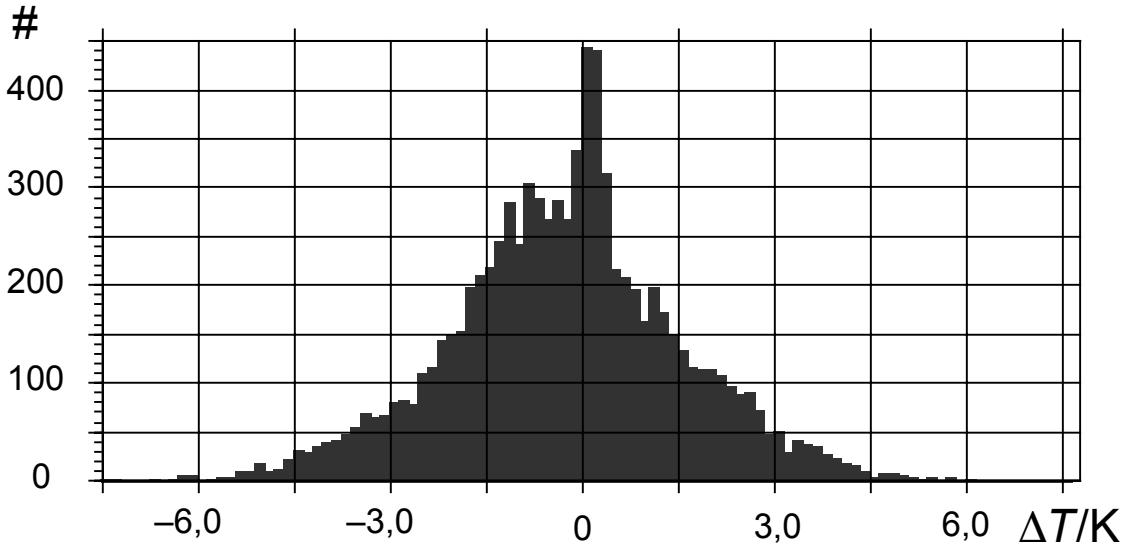

Fig.4. Mixture distributions for the results of CCT key comparisons: K3, melting point of indium ($B = 8000$).

6. INTERCOMPARISON OUTCOMES: THE REFERENCE VALUE [8]

From the above discussion, the need of different statistical treatments for the two classes of IC's arises, to identify the suitable procedure to obtain a proper estimate of the Reference Value (RV) and its uncertainty accordingly to the definition of the involved random variables or quantities.

6.1 RV for Artefact Inter-comparisons

The Reference Value, Y_r , which is a *new stochastic variable* of a *second-rank hierarchical level* that combines the N local summary values. The computation of the RV is commonly performed through the weighted mean. [8]
In the j-th participant laboratory, a summary statistics S_j is adopted to combine the values (x_{j1},…, $x_{jH_{ij}}$) and to estimate the local mean. This estimation corresponds to the *first-rank* hierarchy in the computational treatment of the IC data (repeated measurement approach). The decisions taken to process the samples of each Laboratory include also the choice of the summary statistics, among the many available (mean, weighted mean, median, …)[9] that are commonly used to identify the location parameter of the underlying local probability distribution. Usually, the decision is driven mainly by the consideration of being the best suited for the specific experimental case in a Laboratory. Hence, different participants may choose different statistics, causing S_j to have a different statistical probability distribution.[10] To predict accurate estimates, several sources of variations within and between the populations have to be taken into account.

In the literature on IC studies, the computation of y_r and u_r based on the weighted mean estimator, does not generally take into account the effect of several factors on the quality of the estimates: small size of the IC sample, N (usually < 20), and the peculiarities of F_r and insufficient credibility of some uncertainty estimates. A possible drawback is that the above approach would generally prevent from taking into account a possible bimodal density f_r , even should a visual inspection of the data have suggested that: it should be stressed that, if a density probability function is multi-modal, the choice of a single location (or summary) parameter, y_r, could not be sufficient in representing the specific data variability of the IC. In general, in some conditions the classical approach based on the sampling theory, i.e., on N large, may give a misleading result.

6.2 RV for Physical-state Realisation Inter-comparisons

---

[8] The Reference Value can sometimes be defined as a deterministic parameter (DRV), i.e., as a numerical value, r. A DRV has been preferred, for example, for the temperature key comparisons (KC's) CCT K2 and CCT K4 [11,13]. The DRV value, irrespective to the method used to obtain it, is purely a *stipulated* one, as opposite to a statistically-generated one, and its purpose is limited to the computation of the absolute degree of equivalence –generally required by the MRA– discussed in Section 7, where it determines the offset of the measurand value of every participant.
This kind of RV definition is not in contrast with the MRA requirements («*the key comparison reference value with its associated uncertainty*», which can be zero) and can have the advantage to avoid the consequences of a critical choice of the summary stochastic variable, when its meaning or use is not essential to the use of the IC results. In addition, it avoids the problems arising from the correlation of the RV with the participant values, as no uncertainty is associated to the DRV.

[9] The median can be the preferred choice for the sake of robustness, since it is known that outliers will not heavily affect the accuracy of the uncertainty estimates.

[10] As an example, the choice of the median as summary statistics involves a discrete sampling distribution function, which only asymptotically (for *the sample size* very large) can be assumed to be normal.

According to the approach of Section 5, the definition of the Reference Value is simply the expected value, r, of the mixture density f that takes into account the local variability:

$$r = E_{F(\Lambda)}(X) \quad \text{or} \quad r = \int x \, f(x, \Lambda) \, dx \tag{8}$$

Equation (8) is the first moment of the mixture density. This implies that the RV is simply a *numerical value* and *not* a random variable. In this definition, the RV cannot be viewed as an estimation resulting from a complex measurement process. However, its definition does not appear to be in contrast to the MRA being based on a sound probability model.

The probability distribution F directly represents the variability of the super-population of the IC physical-state samples. See [14] for a complete discussion. No assumptions are required to define F, except those that have been assumed in each Laboratory and are embedded in the local distributions F_j provided to the IC. Therefore, the determination of a summary value from the probability distribution can be done without undertaking any hierarchical step and by avoiding local decision and statistical procedures that could introduce biases (i.e. in-homogeneities due to the choice of different first-rank location parameters).

If the mixture density is completely known (all the Λ values are known) a numerical integration can be performed to obtain the expected value of (8). However, the analytical computation with some mixture density may require not simple mathematics, especially when $N > 2$, the Monte Carlo approach is often applied to overcome these difficulties in the integration of (10). The mixture model is suitable from a computational point of view and can be easily embedded in a Monte Carlo algorithm. An algorithm is given in [4] that automatically generates values and simply computes the output reference value. When some of the local F_j are only partially available an approximation to the local density can be obtained from the information in (1) as in [14].

For the estimate of the uncertainty associated to the RV, being here a purely numerical value, it can be avoided at all, as for the DRV in Note 5, or a suitable treatment should accordingly be identified. In the latter case, for example, the second moment of the mixture probability distribution could be computed to describe the variability of the probability mass with respect to the first moment –the expected value r, however one may question if it should also be assumed to be a suitable parameter to characterise the IC uncertainty suitable to metrological purposes, and then if is should be associated to the degree of equivalence, as required by the MRA. Since this point requires a more detailed analysis a specific work will be deserved to it.

7. INTERCOMPARISON OUTCOMES: ABSOLUTE DEGREE OF EQUIVALENCE

The Absolute Degree of Equivalence (ADE) is the difference between the measurand value of each participant Laboratory and the Reference Value.

7.1 ADE for Artefact Inter-comparisons

The Absolute Degree of Equivalence of the j-th participant, Y_{jr}, is a *new stochastic variable*, combination of two variables of *second-rank hierarchical level*:

$$Y_{j,r} = Y_j - Y_r \tag{9a}$$
$$F_{j,r} = F_j * F_r \tag{9b}$$
$$u_{j,r} = u(y_{j,r}) \tag{9c}$$

Obviously to computation the uncertainty $u_{j,r}$, one must take into account the *correlation* existing between each the two variables. For example, in the case all the F_j are Normal (whence also F_r is N), it comes $\sigma_{jr}^2 = \sigma_r^2 - \sigma_j^2$ [7].

In the case that a DRV is defined (a deterministic parameter *r*, Note 5), for the Absolute Degree of Equivalence of the *j*-th participant, one can compute the difference between the value provided by the *j*-participant, according to the simple calculations recalled in Section 3.1, and the consensus value *r*:

$$y_{j,r} = y_j - r \qquad (10, a)$$

However, in this case the stochastic variability can be described only by the local pdf, and then it can be assumed the following:

$$F_{j,r} = F_j \qquad (10, b)$$
$$u_{j,r} = u(y_j) \qquad (10, c)$$

7.2 ADE for Physical-state Realisations Inter-comparisons

The Absolute Degree of Equivalence of the *j*-th participant, is now the *numerical value*

$$a_{j,r} \equiv E(F_j) - r \; . \qquad (11)$$

Also in this case the association of an uncertainty will be driven after a deeper discussion on the general definition of «uncertainty» in IC framework, which it is still in progress in the metrological community.

8. CONCLUSIONS

This paper discussed a classification of the inter-comparisons that is relevant to identify the proper statistical method to combine the data provided for each participant to the inter-comparison. The proposed approach for Class 2 IC's constructs a single probabilistic model for the reference distribution probability function, based on the use of the mixture density model. This approach allows the estimate of the reference value simply as the expected value of the mixture density function. The method does not require strong assumptions –as *N* large– or limitations on the local probability distributions, such as the Normality density in each participant laboratory. It is particularly valuable for inter-comparisons of physical-state realisations, e.g., for temperature standards, where the population associated to an IC can be viewed as a super-population. However, Class 2 inter-comparisons probably include a wider range of IC's, such as the ones where a single standard is circulated and measured.
The paper also compared the main features of the statistical treatments suitable for the IC outcomes in both cases of Class 1 and Class 2.
Some aspects of the treated problems, especially the meaning of uncertainty in Class 2 IC's and the consequent approach, still deserve a deeper subsequent insight.

ACKNOWLEDGEMENTS
The author wishes to acknowledge the valuable discussions with Giuseppe Regoliosi.

REFERENCES

1. 1. Ciarlini P., *Advantages of the bootstrap technique in solving inter-laboratory comparison problems*. NPL Seminar «Statistical Analysis of Inter-laboratory Comparisons», Teddington, November 1999.
2. Ciarlini P., Regoliosi G, Pavese F. In *Advanced Mathematical Tools in Metrology* , Ciarlini, Cox, Monaco, Pavese Eds, Series on Advances in Mathematics for Applied Sciences Vol.**16**,World Scientific, Singapore, 219-229.
3. Pavese F., *Advanced Mathematical Tools in Metrology*, vol.4, P.Ciarlini, A.B.Forbes, F.Pavese and D.Richter eds., *Series on Advances in Mathematics for Applied Sciences* **53**, World Scientific, Singapore, 2000, 182-196.
4. Ciarlini P., Pavese F., Regoliosi G., *A bootstrap algorithm for mixture models and interval data in intercomparisons*. Conference «Algorithms for Approximation IV», Huddersfield, July 2001, Proceedings 2002, 138-145.
5. Ciarlini P., Pavese F., *Some basic questions in the treatment of intercomparison data*, VNIIM-NPL Workshop, St.Petersburg, June 2002; *Izmeriskaya Tecnichka*, 2002, in press.
6. Davison A.C.. *Bootstrap methods, in Supplement to Handbook of Applicable Mathematics* , W. Ledermann. E. Lloyd. S. Vajda and C.Alexander (eds.), John Wiley, New York ,1988.
7. Ciarlini P., Regoliosi G., Pavese F., Advanced Mathematical Tools in Metrology, vol.1, P.Ciarlini, M.G.Cox, R.Monaco and F.Pavese eds., Series on Advances in Mathematics for Applied Sciences **16**, World Scientific, Singapore, 1994, 219-230.
8. Cox M.G., *Metrologia*, 2002, **39**, 587-588, 589-596.
9. McLachlan G.J., Basford K.E., *Mixture models: inference and applications to clustering*, Dekker Pub., N.Y. (1988).
10. Steele A. G., Hill K. D., Douglas R. J., *Metrologia* 2002, **39**, 269-278.
11. Final Report of Key Comparisons CCT-K2, Database of Appendix B of MRA (www.bipm.org), BIPM, Sèvres, 2002. Summary: Steele A. G., Fellmuth B., Head D. I., Hermier Y., Kang K. H., Steur P. P. M., Tew W. L., *Metrologia*, 2002, **39**, 551-572.
12. Final Report of Key Comparisons CCT-K3, Database of Appendix B of MRA (www.bipm.org), BIPM, Sèvres, 2002. Summary: Mangum B. W., Strouse G. F., Guthrie W. F., Pello R., Stock M., Renaot E., Hermier Y., Bonnier G., Marcarino P., Gam K.S., Kang K. H., Kim Y., G., Nicholas J. V., White D. R., Dransfield T. D., Duan Y., Qu Y., Connolly J., Rusby R. L., Gray J., Sutton G. J. M., Head D. I., Hill K. D., Steele A., Nara K., Tegeler E., Noatsch U., Heyer D., Fellmuth B., Thiele-Krivoj B., Duris S., Pokhodun A. I., Moiseeva N. P., Ivanova A. G., de Groot M. J., Dubbeldam J. F., *Metrologia*, 2002, **39**, 179-207.
13. Final Report of Key Comparisons CCT-K4, Database of Appendix B of MRA (www.bipm.org), BIPM, Sèvres, 2002. Summary: Nubbemeyer H.G., Fischer J., *Metrologia*, 2002, **39**, Tech. Suppl., 03001.
14. Ciarlini P., Cox M.G., Pavese F., Regoliosi G., in preparation.